Statistics Canada Statistique Canada

Methodology Branch

Census & Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes de recensement
et d'enquêtes ménages

Canada

S-185E
c.2

ON SMOOTHED ESTIMATES OF UNEMPLOYMENT RATES FROM LABOUR FORCE SURVEY DATA

S. Kumar and J.N.K. Rao

Statistics Canada and Carleton University

Methodology Branch Working Paper

Number: CHSM 85-079E

ON SMOOTHED ESTIMATES OF UNEMPLOYMENT RATES FROM LABOUR FORCE SURVEY DATA

S. Kumar and J.N.K. Rao

Statistics Canada and Carleton University

Smoothed estimates of unemployment rates are obtained by fitting a model to survey estimates of cell proportions in a two-way table. The smoothed estimates are shown to be considerably more efficient than the survey estimates, especially for cells with small samples.

1. INTRODUCTION

The Canadian Labour Force Survey (LFS) provides precise estimates of monthly unemployment rates at the national level and also at the regional levels. However, the survey estimates of unemployment rates in cross-classifications are less precise and the coefficient of variation (CV) for cells with small samples could be unacceptably high. By fitting a model to survey estimates of cell proportions, it is possible to obtain smoothed estimates which are considerably more efficient than the survey estimates.

In this note, smoothed estimates under a logistic regression model are given and the associated standard errors are obtained. An application to data from the October 1980 LFS is also given. The empirical results indicate that the smoothed estimates are considerably more efficient than the survey estimates, especially for cells with small samples.

2. SMOOTHED ESTIMATES

Suppose that the population of interest is partitioned into I domains (cells) according to the levels of one or more factors. Let $\hat{N}_i$ denote the survey estimate of i-th domain size N_i $(i=1, .., I; \Sigma N_i = N)$. The corresponding estimate of i-th domain total, N_{i1}, of a binary (0, 1) response variable is denoted by $\hat{N}_{i1}$. The ratio estimate $\hat{p}_i = \hat{N}_{i1}/\hat{N}_i$ is used to estimate the cell proportion $p_i = N_{i1}/N_i$.

We fit a logistic regression (logit) model to the cell proportions p_i given by $p_i = f_i(\underset{\sim}{\beta})$, where

$$\ln\left[\frac{f_i}{1-f_i}\right] = \text{logit } f_i = \underset{\sim}{x}_i'\underset{\sim}{\beta}, \quad i = 1, .., I \tag{1}$$

In (1), $\underset{\sim}{x}_i$ is an s-vector of known constants obtained from the factor levels and $\underset{\sim}{\beta}$ is the s-vector of unknown parameters. The smoothed estimates, $\hat{f}_i$, of p_i under model (1) are obtained from "pseudo" likelihood equations given by

$$X'D(\underset{\sim}{w})\hat{\underset{\sim}{f}} = X'D(\underset{\sim}{w})\hat{\underset{\sim}{p}}. \tag{2}$$

Here $\hat{\underset{\sim}{f}} = (\hat{f}_1, .., \hat{f}_I)'$, $\hat{f}_i = f_i(\hat{\underset{\sim}{\beta}})$, $\hat{\underset{\sim}{p}} = (\hat{p}_1, .., \hat{p}_I)'$, $D(\underset{\sim}{w}) = \text{diag}(w_1, .., w_I)$, $w_i = N_i/N$ where $N = \Sigma N_i$, and $X' = (\underset{\sim}{x}_1, .., \underset{\sim}{x}_I)$. The solutions $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{f}}$ of (2) are obtained by iterative calculations.

Let $\hat{V}$ denote the estimated covariance matrix of the survey estimates $\hat{\underset{\sim}{p}}$. Then, the corresponding estimated covariance matrix of the smoothed estimates, $\hat{\underset{\sim}{f}}$, is given by (Roberts, 1984).

$$\hat{V}_f = D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1}-\hat{\underset{\sim}{f}}) \; X \; \hat{V}_\beta X'D(\hat{\underset{\sim}{f}})D(\underset{\sim}{1}-\hat{\underset{\sim}{f}}), \tag{3}$$

where $\hat{V}_\beta$ is the estimated covariance matrix of $\hat{\underset{\sim}{\beta}}$. Here $D(\hat{\underset{\sim}{f}}) = \text{diag}(\hat{f}_1, .., \hat{f}_I)$, $D(\underset{\sim}{1}-\hat{\underset{\sim}{f}}) = \text{diag}(1-\hat{f}_1, .., 1-\hat{f}_I)$ and

$$\hat{V}_\beta = (X'\hat{\Delta}X)^{-1}(X'D(\underset{\sim}{w})\hat{V}D(\underset{\sim}{w})X)(X'\hat{\Delta}X)^{-1}, \tag{4}$$

where

$$\hat{\Delta} = \text{diag}(w_1\hat{f}_1(1-\hat{f}_1), \ldots, w_I\hat{f}_I(1-\hat{f}_I)).$$

(See also Binder, 1983).

The diagonal elements, $\hat{v}_{ii}(f)$, of $\hat{V}_f$ give the estimated variances of smoothed estimates $\hat{f}_i$, while the diagonal elements $\hat{v}_{ii}$, of $\hat{V}$ are the estimated variances of survey estimates $\hat{p}_i$.

A scalar measure of efficiency of smoothed estimates over survey estimates is given by

$$e = \frac{tr(\hat{V})}{tr(\hat{V}_f)} \quad \frac{\sum_i \hat{v}_{ii}}{\sum_i \hat{v}_{ii}(f)}, \qquad (5)$$

where tr denotes the trace operator.

Kumar and Rao (1984) provide a test of goodness-of-fit of the logit model (1), taking account of the survey design. If the model provides an adequate fit, then the bias of smoothed estimates should be small relative to standard error. The smoothed estimates, $\hat{f}$, are similar to synthetic estimators used in small area estimation.

3. APPLICATION TO LFS DATA

Kumar and Rao (1984) fitted a logit model to data from the October 1980 LFS. The sample consisted of males aged 15-64 who were in labour force and not full-time students. Two factors, age and education, were employed to explain the variation in unemployment rates via a logit model. Age groups were formed by first dividing the interval [15, 64] into ten groups [10 + 5j, 14 + 5j], j = 1, ..., 10 and then using the mid-point of each interval, A_j, as the value of age for all persons in that age group. Similarly, the levels of education E_k, were formed by assigning to each person a value based on median years of schooling, resulting in six levels: 7, 10, 12, 13, 14 and 16. Thus the age by education classification provided a two-way table of I = 60 survey estimates $\hat{p}_{jk}$ of employment rates p_{jk}. The survey estimates $1-\hat{p}_{jk}$ of unemployment rates and their estimated CV's are given in Table 1, where the cells are lexicographically ordered. The CV's for two cells (numbered 6 and 54) with $1-\hat{p}_{jk} = 0$ are undefined.

Kumar and Rao (1984) have shown that the following simple logit model provides an adequate fit to the data:

$$\ln \frac{f_{jk}}{1-f_{jk}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_k. \qquad (6)$$

The estimates of β_i are given by

$$\hat{\beta}_0 = -3.10,\ \hat{\beta}_1 = 0.211,\ \hat{\beta}_2 = -0.00218,\ \hat{\beta}_3 = 0.1509.$$

The corresponding smoothed estimates $1-\hat{f}_{jk}$ are given in Table 1 along with their CV's. It may be noted that the smoothed estimates remain non-zero even when the corresponding survey estimates are zero. Their CV's are also well-defined for all the cells.

It is clear from Table 1 that the CV's of survey estimates are quite large, especially for cells with small samples. The CV's range from 6.8% (for cell 3) to 98.5% (for cell 59). On the other hand, the smoothed estimates $1-\hat{f}_{jk}$ lead to dramatic reductions in CV. The CV of smoothed estimates range from 3.3% (cell 8) to 12.4% (cell 60); the CV for cell 59 is reduced from 98.5% to 11.0%. The average CV of $1-\hat{p}_{jk}$ (over the 58 cells with $1-\hat{p}_{jk} > 0$) is 32.1% compared to 6.2%, the average CV of $1-\hat{f}_{jk}$ over all the 60 cells.

The efficiency measure e gives

$$e = \frac{0.01597}{0.00089} = 17.9,$$

showing that the smoothed estimates lead to large gains in efficiency over the survey estimates. Moreover, the bias of smoothed estimates should be relatively small since the model (6) provides an adequate fit to the data.

4. ADJUSTED SMOOTHED ESTIMATES

Sometimes it is desirable to adjust the smoothed estimates, $\hat{f}_i$, so that the estimate at an aggregate level is identical to the corresponding survey estimate with a small C.V. For instance, in the LFS application, we may want $\Sigma_k w_{jk} \hat{f}_{jk} = \Sigma_k w_{jk} \hat{p}_{jk}$ for each j, where $w_{jk} = \hat{N}_{jk}/\hat{N}$ is the survey estimate of the population proportion in (j,k)-th cell.

Suppose the $\hat{f}_i$ are to be adjusted over a set, A, of cells so that the resulting aggregate estimate $\tilde{f}_A$ is identical to the aggregate survey estimate $\hat{p}_A = \sum_{i \in A} w_i \hat{p}_i$, i.e., we wish to find $\tilde{f}_i$, $i \in A$ such that $\tilde{f}_A = \sum_{i \in A} w_i \tilde{f}_i = \sum_{i \in A} w_i \hat{p}_i = \hat{p}_A$. This can be easily accomplished by adding the bias correction $\sum_{i \in A} w_i (\hat{p}_i - \hat{f}_i) / \sum_{i \in A} w_i$ to each $\hat{f}_i$, $i \in A$. Thus the adjusted estimates are given by

$$\tilde{f}_i = \hat{f}_i + \left[\sum_{i \in A} w_i (\hat{p}_i - \hat{f}_i) / \sum_{i \in A} w_i \right], \quad i \in A \tag{7}$$

Battese and Fuller (1982) proposed this simple adjustment procedure in the context of small area estimation, but the weights w_i in their application are constant.

Alternatively, a simple ratio adjustment gives

$$\tilde{f}_i(r) = \left(\sum_{i \in A} w_i \hat{p}_i / \sum_{i \in A} w_i \hat{f}_i \right) \hat{f}_i, \quad i \in A. \tag{8}$$

Since w_i are random variables in the LFS application, the evaluation of variance of $\tilde{f}_i$ or $\tilde{f}_i(r)$ is somewhat complicated, but can be done by the familiar linearization method.

REFERENCES

[1] Battese, G.E. and Fuller, W.A. (1982). An error components model for prediction of county crop areas using survey and satellite data. Technical Report, Iowa State University.

[2] Binder, D.A. (1983). On the variances of asymptotically normal estimates from complex surveys. Intl. Statist. Review, 51, pp. 279-292.

[3] Kumar, S. and Rao, J.N.K. (1984). Logistic regression analysis of labour force survey data. Survey Methodology, 10, pp. 62-81.

[4] Roberts, G. (1984). On chisquared tests for logit models with cell proportions estimates from survey data. Unpublished manuscript. Carleton University.

TABLE 1: Survey estimates $1-\hat{p}_{jk}$ and smoothed estimates $1-\hat{f}_{jk}$ of unemployment rates and associated CV's in percent

Cell No.	$1-\hat{p}_{jk}$	$CV(1-\hat{p}_{jk}) \times 100$	$1-\hat{f}_{jk}$	$CV(1-\hat{f}_{jk}) \times 100$
1	.268	12.7	.286	5.1
2	.238	6.8	.203	4.4
3	.132	10.0	.159	4.7
4	.175	26.9	.140	5.1
5	.117	58.3	.122	5.7
6	0	-	.094	7.2
7	.112	15.8	.176	4.8
8	.128	8.8	.120	3.3
9	.090	8.9	.091	3.4
10	.097	16.5	.080	3.9
11	.061	19.7	.069	4.7
12	.063	29.8	.052	6.5
13	.094	16.4	.113	5.1
14	.081	11.4	.075	3.5
15	.047	13.2	.056	3.6
16	.054	19.6	.049	4.1
17	.047	20.3	.042	4.9
18	.032	23.7	.032	6.7
19	.095	17.3	.078	5.7
20	.052	14.6	.051	4.2
21	.039	16.9	.038	4.4
22	.030	30.5	.033	4.9
23	.029	26.0	.029	5.6
24	.019	25.6	.021	7.3
25	.062	15.1	.059	6.0
26	.030	20.9	.038	4.7
27	.015	28.5	.029	5.0
28	.041	44.1	.025	5.5
29	.031	29.8	.021	6.1
30	.014	40.3	.016	7.8

TABLE 1: Survey estimates $1-\hat{p}_{jk}$ and smoothed estimates $1-\hat{f}_{jk}$ of unemployment rates and associated CV's in percent (Cont'd)

Cell No.	$1-\hat{p}_{jk}$	$CV(1-\hat{p}_{jk}) \times 100$	$1-\hat{f}_{jk}$	$CV(1-\hat{f}_{jk}) \times 100$
31	.047	18.4	.049	6.0
32	.027	18.6	.032	4.8
33	.027	33.2	.024	5.2
34	.017	60.4	.020	5.7
35	.029	39.4	.018	6.4
36	.0076	53.8	.013	8.1
37	.048	16.1	.045	5.8
38	.024	24.1	.029	4.7
39	.037	30.5	.022	5.2
40	.033	49.7	.019	5.8
41	.023	40.6	.016	6.5
42	.0062	80.3	.012	8.3
43	.054	14.4	.046	5.8
44	.028	30.3	.030	5.1
45	.031	35.2	.022	5.7
46	.018	73.3	.019	6.3
47	.023	53.9	.017	7.0
48	.010	67.8	.012	8.8
49	.046	16.7	.053	6.9
50	.031	25.2	.034	6.6
51	.012	41.1	.025	7.2
52	.011	49.1	.022	7.8
53	.024	52.0	.019	8.4
54	0	-	.014	10.0
55	.069	15.5	.066	9.3
56	.040	25.6	.043	9.3
57	.038	39.6	.032	10.0
58	.080	53.1	.028	10.5
59	.0069	98.5	.024	11.0
60	.019	76.1	.018	12.4